Lecture #24

- Undergirding much of what we have done is based on interpretting data as generated as iid samples from an unknown distribution, perhaps a member of a known parametric family: given data $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$, estimate θ according to some methodology yielding $\hat{\theta}$.

- We might ask that $\hat{\theta}$ has some properties like:
  (i) On average, $\hat{\theta}$ gets it right: $\mathbb{E}(\hat{\theta}) = \theta$, i.e. $\hat{\theta}$ is unbiased
  (ii) $\hat{\theta}$ doesn't deviate too much across samples: $\mathrm{Var}(\hat{\theta})$ is small
  (iii) $\hat{\theta}$ converges to the correct thing in the large sample limit: $\lim_{n\to\infty} \hat{\theta} - \theta$, i.e. $\hat{\theta}$ is consistent.

- In this sense, we have been interested in how the random quantity $(\hat{\theta})$ relates to the fixed but unknown parameter(s) (θ).

- This basic world view thinks of probability as fundamentally related to long-run frequencies: $\mathbb{P}(X=0) = \frac{1}{2}$ means if I sample X ∞-many times, half the time I get $X=0$." This is basically the LLN.

- Such a view of statistics is called frequentism. We may think of it as a world view with three basic postulates about how one can discuss random quantities:
  (1) Probabilities are limiting / long run frequencies, and are therefore objective (but unknown)
  (2) When doing parametric modeling, the parameters are fixed and deterministic.
  (3) The only reasonable way to design statistical procedures is to ensure that their long-run frequency properties are controlled.

- This is the idea of a confidence interval: a collection of ∞-many random confidence intervals should contain the true parameter a fixed proportion of of the time. The interval is random, not the parameter of interest.

- This is very nice mathematically, because the tools of probability theory and analysis become highly useful.

- It is, however, limiting. One can only do statistics insofar as they refer to long-run averages. An example from Wasserman is instructive: "$\mathbb{P}$(Einstein drank tea on 8-1-1948) = .35" makes no sense from a frequentist standpoint, because the event "Einstein drank tea on 8-1-1948" does not refer to any long-run probabilities, taken literally.

- But of course, the above statement does make sense to most humans! What it means is, with everything we know about the world, we feel somewhat confident (but not very much) that Einstein drank tea on that day.

- In normal language, probabilities can be used to convey measure of confidence / degree of belief, in addition to long-run frequency. This insight/world view is captured by the perspective of Bayesian statistics.

- In contrast with the tenets of frequentism, we might list those of Bayesianism as:

(1.) Probabilities are degrees of belief, and therefore subjective.

(2.) When doing parametric inference, we may make probabilistic statements about parameters depending on our (changing) degree of belief.

(3.) We may make inferences about parameters by estimating distributions over parameters, which may not relate to any notion of long-term frequency.

· The way one utilizes this world view in practice gives the method its name: Bayes Theorem is central.

· Let us recall: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Applying this in the context of $A = \{X = x\}$, i.e. we observe data $X = x$

$B = \{\theta = \theta^*\}$, i.e. the underlying parameter takes value θ^*,

we get $$\mathbb{P}(\theta = \theta^* \mid X = x) = \frac{\mathbb{P}(X = x \text{ and } \theta = \theta^*)}{\mathbb{P}(X = x)}$$

$$= \frac{\mathbb{P}(X = x \mid \theta = \theta^*) \cdot \mathbb{P}(\theta = \theta^*)}{\sum_{\theta_0} \mathbb{P}(X = x \mid \theta = \theta_0) \cdot \mathbb{P}(\theta = \theta_0)}$$

- In the continuous context, everything is done in terms of densities and we get

$$\underbrace{f(\theta^* \mid x)}_{\text{what I want}} = \frac{\overbrace{f(x \mid \theta^*)}^{\text{easy}} f(\theta^*)}{\underbrace{\int_{\mathbb{R}} f(x \mid \theta) f(\theta)\, d\theta}_{\text{???}}} \quad \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \text{Big idea of Bayesian statistics.}$$

· Notice that (1) lets us turn statements about how parameters are determined by data ($f(\theta^* \mid x)$) into statements about how data depend on parameters ($f(x \mid \theta^*)$). The latter is very easy to compute in a parametric setting.

· We do, however need to deal with the other two factors on the RHS:

(i) $f(\theta^*)$: we need a latent distribution over the parameters. This is something the user needs to bring to the table. We call it a prior distribution, since data doesn't inform it. We need prior knowledge of its structure

(ii) $\int_{\mathbb{R}} f(x|\theta) f(\theta) d\theta$: This is basically a normalizing constant, we will, in general, have a hard time computing it. But, in many cases we don't need it!

- Notice that if we are thinking of θ^* as something to estimate, then $f(x|\theta^*)$ is the likelihood function! So,

$$f(\theta^*|x) = \frac{f(x|\theta^*) f(\theta^*)}{\int_{\mathbb{R}} f(x|\theta) f(\theta) d\theta}$$

$$= \frac{\mathcal{L}(\theta^*|x) f(\theta^*)}{\text{normalizing constant}}$$

$$\propto \mathcal{L}(\theta^*|x) f(\theta^*)$$

($\propto$: proportional to)

- This captures the practical interpretation of Bayesian statistics:

$$\underbrace{f(\theta^*|x)}_{\text{"posterior" distribution}} \propto \underbrace{\mathcal{L}(\theta^*)}_{\text{likelihood}} \underbrace{f(\theta^*)}_{\text{prior}}$$

"posterior" distribution: distribution over parameter space having observed data x

likelihood × prior: adjust our initial understand of the parameter distribution using some a data-driven update factor, i.e. the likelihood function

* : Suppose $X_1, \dots, X_n \sim \text{Bernoulli}(\rho)$. What is the posterior distribution over ρ (the parameter) given $\{x_i\}_{i=1}^n$ (the data)? Well, it depends on our prior distribution! Let's try something that seems easy. $f(\rho) = \begin{cases} 1, & \rho \in [0,1] \\ 0, & \text{else} \end{cases}$

i.e. a uniform prior over $[0,1]$.

Then $$f(\rho^* \mid x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n \mid \rho^*) \cdot f(\rho^*)}{\int_{\mathbb{R}} f(x_1, \dots, x_n \mid \rho) f(\rho)\, d\rho}$$

$$= \frac{\mathcal{L}(\rho^*) \cdot 1}{\int_{\mathbb{R}} \mathcal{L}(\rho)\, d\rho}$$

$$\propto (1-\rho^*)^{1-\sum_{i=1}^n x_i} \cdot [\rho^*]^{\sum_{i=1}^n x_i}$$

Letting $s := \sum_{i=1}^n x_i$, this gives,

$$f(\rho^* \mid x_1, \dots, x_n) \propto (1-\rho^*)^{1-s} \cdot [\rho^*]^{s}.$$

As a r.v. over ρ^*, this is a β-distribution with parameters $(s+1)$ and $(n-s+1)$, i.e. $\rho \mid \{x_i\}_{i=1}^n \sim \text{Beta}\left(\sum_{i=1}^n x_i + 1,\ n - \sum_{i=1}^n x_i + 1\right)$.

ρ is a r.v. impacted by observations $\{x_i\}_{i=1}^n$